The 30th European Conference on Information Retrieval (ECIR 2008)
Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'08)

Web Search Disambiguation by Collaborative Tagging

Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt

UNIVERSITY OF Southampton
School of Electronics and Computer Science

30th March 2008

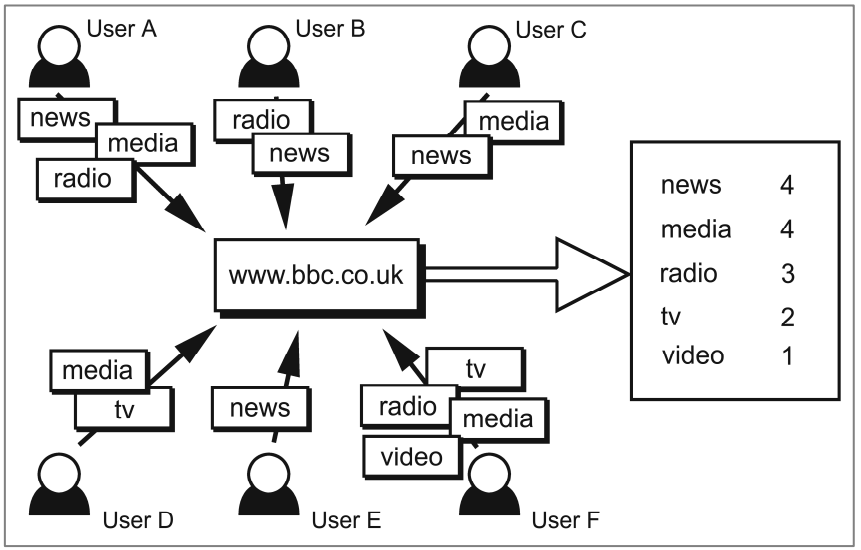
Introduction

- **Problems of Web Search**
  - Queries by ambiguous terms return many irrelevant results
  - **Example**: *bridge*

    Search results contain pages about bridge as:
    1) a kind of card games;
    2) a form of architectural structure;
    3) a design pattern in software development;
    4) a device in computer networking
  - **Problem**
    (1) Low precision and recall
    (2) Users need to filter the results by themselves

Introduction

- **Collaborative Tagging Systems**
  - Very popular (e.g. del.icio.us, Flickr, Bibsonomy)
  - Aggregate user-contributed metadata of resources
  - Provide rich information about the relations between different tags
  - Sources for understanding how keywords are used
  - An Example: →



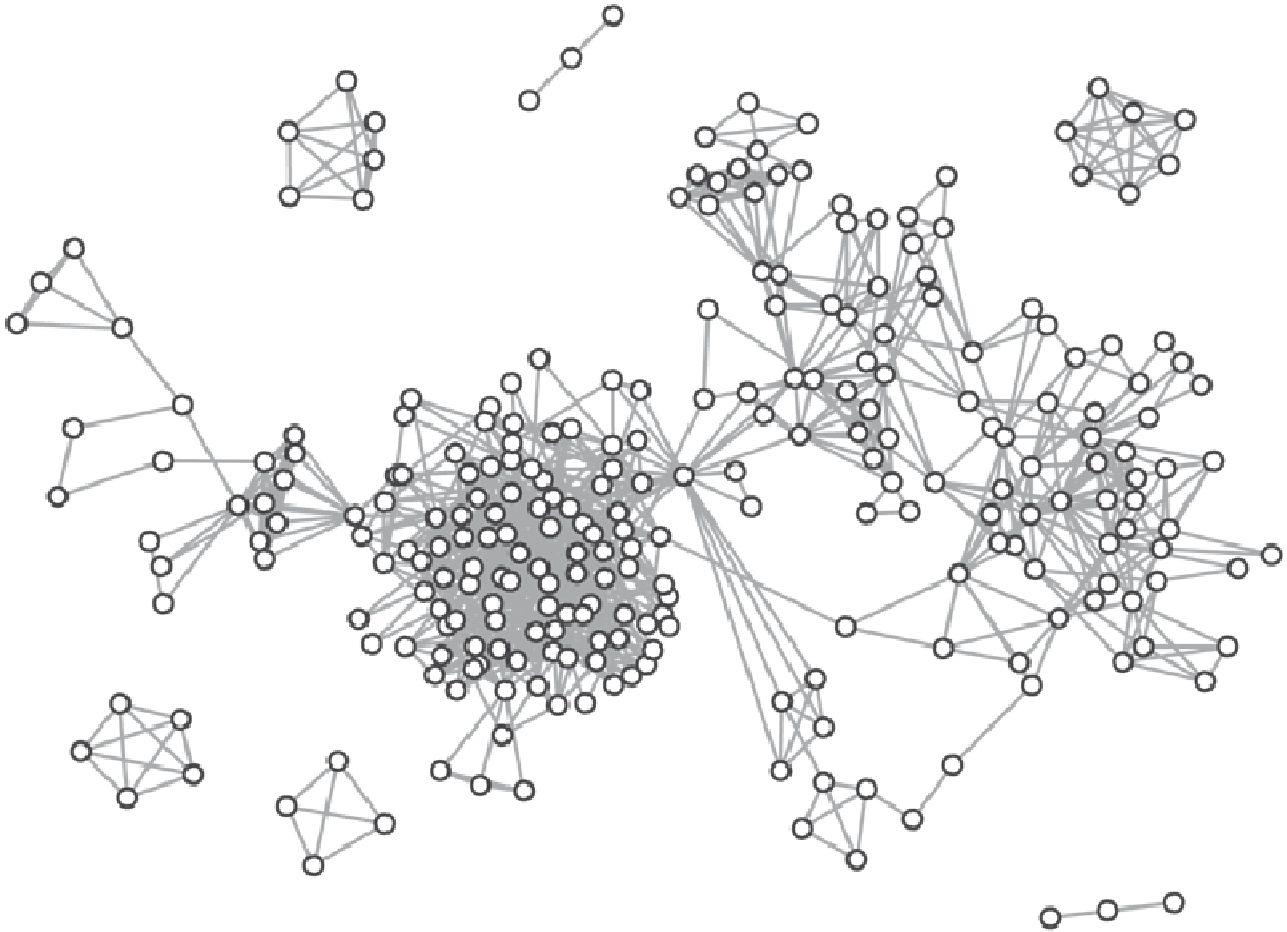
Introduction

- **Our Proposal**
  - Making use of the information available in collaborative tagging to enhance Web search
  - Step 1:
    Discover the different contexts in which tags are used in collaborative tagging by clustering
  - Step 2:
    Apply the results in the form of sets of tags to classify search results returned by search engines

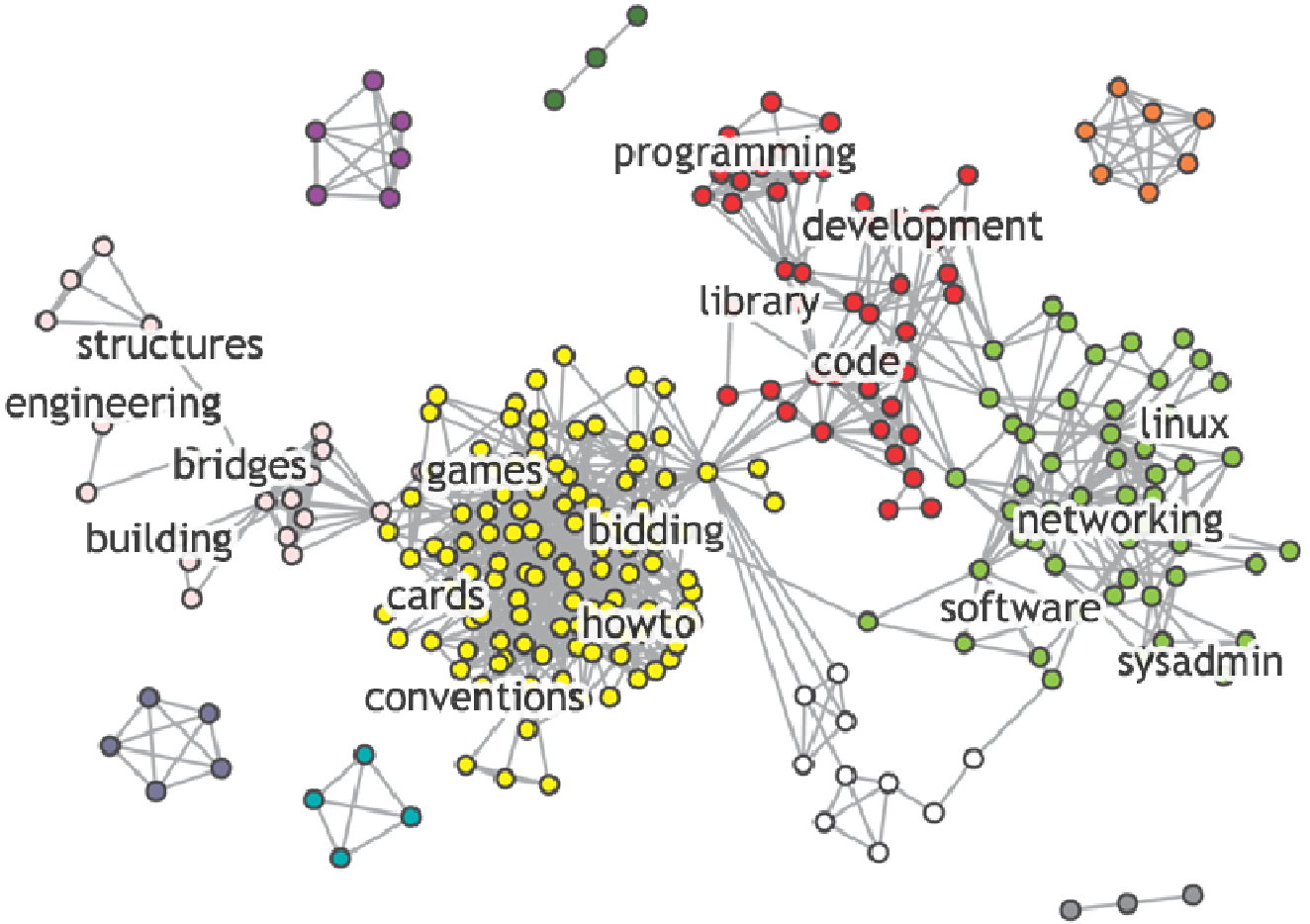
Tag Meanings

- Most users in del.icio.us use a tag for the same meaning for most of the time (e.g. *sf*)
- It implies that clustering technique can be used to identify the different contexts
- Proposed Algorithm:
  1. Construct a document network from a folksonomy
  2. Cluster documents based on the users who have used the tag on the documents
  3. Extract frequently co-occurred tags as representations of the contexts

Tag Meanings



Tag Meanings



Tag Meanings

- Contexts in which *bridge* is used:

Design pattern	bridge, programming, development, library, code, ruby, tools, software, adobe, dev
Card game	bridge, games, cards, game, imported, howto, conventions, card, bidding, online
Computer networking	bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security
Architecture	bridge, bridges, structures, engineering, science, physics, school, education, building, reference

Web Search Disambiguation

- $\{T_{t,1}, T_{t,2}, \ldots, T_{t,n}\}$ be the set of sets of tags discovered by the tag disambiguation process
- $\mathbf{D_t}$: the set of documents returned by a search engine given the query t
- $K_{t,j}$: the set of keywords characterising a document d_j in the set $\mathbf{D_t}$
- Compare $K_{t,j}$ with each of the $T_{t,i}$'s, determine which category should the document be classified:

$$match(K_{t,j}, T_{t,i}) = \frac{|K_{t,j} \cap T_{t,j}|}{|T_{t,i}|}$$

$$Cat_A(d_j, t) = \begin{cases} \underset{i}{\text{argmax}}\ match(K_{t,j}, T_{t,i}), & \text{if } \max_i match(K_{t,j}, T_{t,i}) \geq \beta \\ 0, & \text{if } \max_i match(K_{t,j}, T_{t,i}) < \beta \end{cases}$$

Evaluation

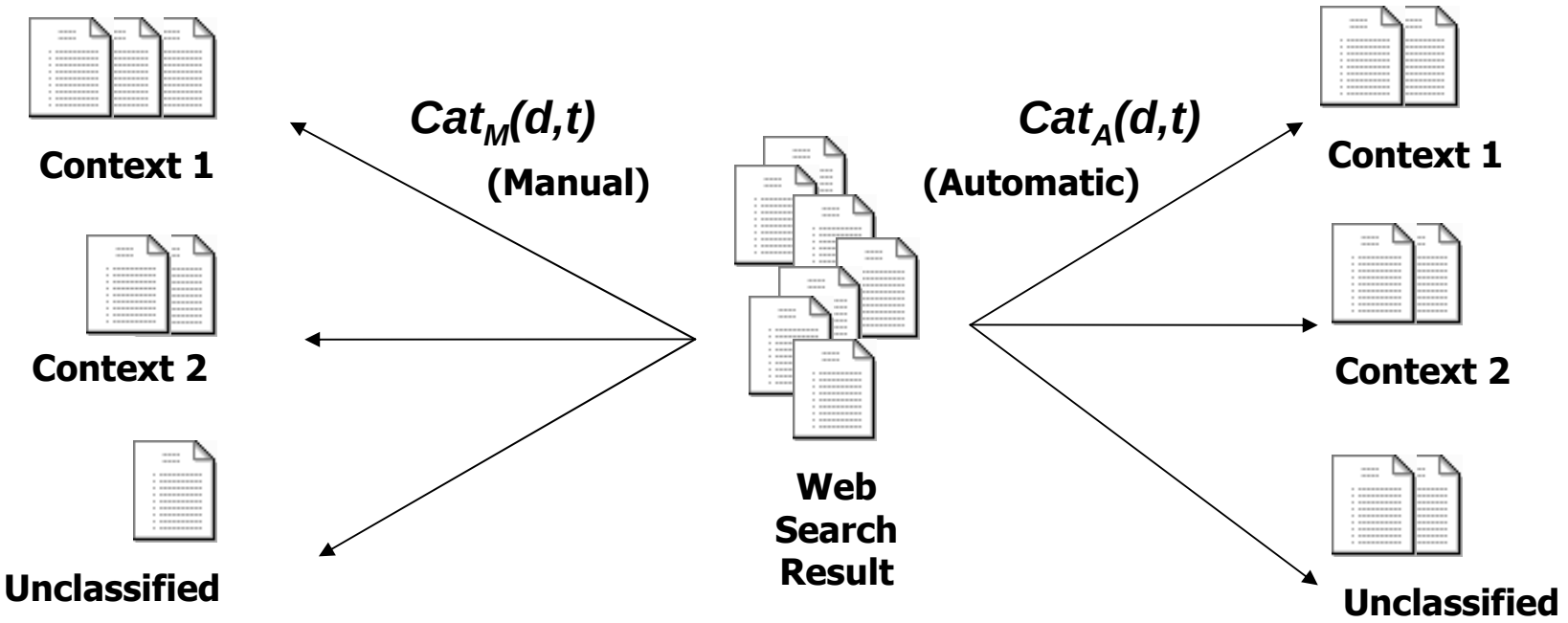
- Experimenting the algorithm on four tags:
  sf, tube, bridge, wine
- Data from del.icio.us collected and the proposed algorithm applied
- Dataset:
  - **del.icio.us**: 50 items tagged by the greatest number of users with the tag in question
  - **Google**: First 50 documents returned by the a query constructed from the tag in question
- A set of keywords is constructed to represent each document returned

Evaluation

Tag	Context	Tags Extracted
sf	San Francisco	sf, sanfrancisco, bayarea, san, francisco, california, travel, events, art, san_francisco
	Science Fiction	sf, scifi, fiction, books, sci-fi, literature, writing, sciencefiction, science, fantasy
tube	YouTube	tube, youtube, video, funny, videos, fun, cool, music, feel.good, flash
	Vacuum Tubes	tube, audio, electronics, diy, amplifier, amp, tubes, music, elect, guitar
	London Underground	tube, london, underground, travel, transport, maps, map, uk, subway, reference
bridge	Design Pattern	bridge, programming, development, library, code, ruby, tools, software, adobe, dev
	Card Game	bridge, games, cards, game, imported, howto, conventions, card, bidding, online
	Computer Networking	bridge, networking, lunux, network, howto, software, sysadmin, firewall, virtualization, security
	Architecture	bridge, bridges, structures, engineering, science, physics, school, education, building, reference
wine	Linux Software	wine, linux, ubuntu, howto, windows, software, tutorial, emulation, reference, games
	Beverage	wine, food, shopping, drink, reference, vino, cooking, alcohol, blog, news

Evaluation

- Documents are first manually classified, represented by the function $Cat_M(d,t)$
- Evaluation involves comparing the classification of the proposed algorithm $Cat_A(d,t)$ with $Cat_M(d,t)$



Evaluation

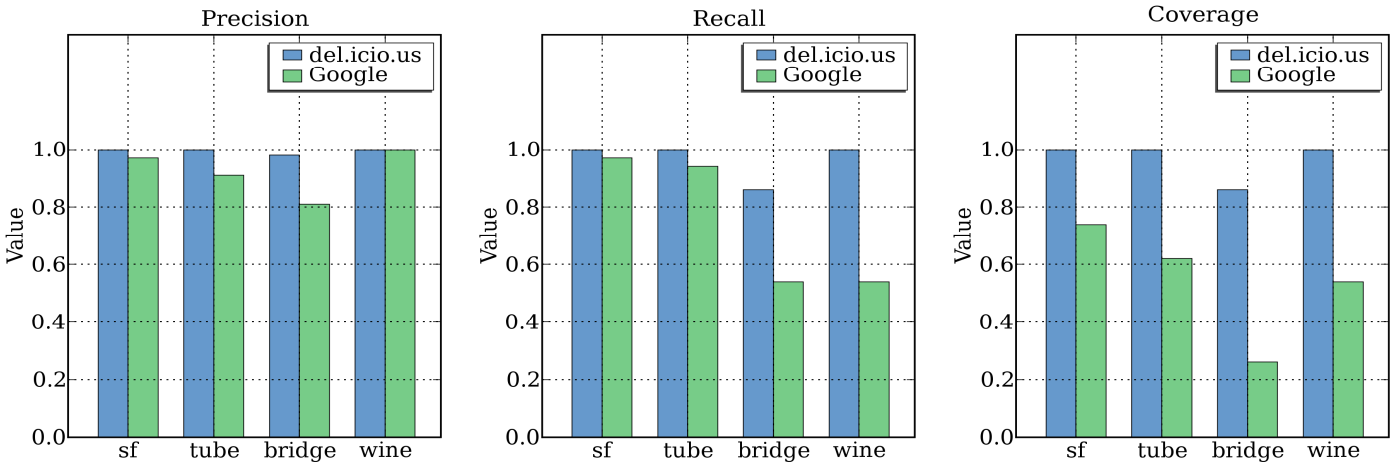
- We employ the following three performance measures:

$$\text{Precision} = \frac{|\{d \in R_t | Cat_M(d,t) = Cat_A(d,t) \wedge Cat_M(d,t) \neq 0\}|}{|\{d \in R_t | Cat_A(d,t) \neq 0\}|}$$

$$\text{Recall} = \frac{|\{d \in R_t | Cat_M(d,t) = Cat_A(d,t) \wedge Cat_M(d,t) \neq 0\}|}{|\{d \in R_t | Cat_M(d,t) \neq 0\}|}$$

$$\text{Coverage} = \frac{|\{d \in R_t | Cat_M(d,t) = Cat_A(d,t) \wedge Cat_M(d,t) \neq 0\}|}{|R_t|}$$

Results and Discussions



- Relatively low recall for Google (54%-97%)
  1. Topics of documents are more diverse
  2. Tags do not match keywords of documents
     (e.g. river, architecture vs. building, engineering)

- Relatively low coverage for Google (26%-74%)
  1. Not all meanings of a tag are identified
     (restricted to how the tag is used in del.icio.us)
  2. Documents are not related to the tag semantically
     (e.g. **B**uilding **R**adio Frequency **ID**entification for the **G**lobal **E**nvironment)

Conclusions and Future Work

- Folksonomies offer rich information on the relations and semantics of tags, and can be used to enhance Web search
- Our proposed method has advantages over use of dictionaries or thesauruses (able to keep up with new meanings)
- Research directions:
  - Increase the comprehensiveness of the sets of tags
  - Identify more contexts in which a tag can be used
  - Better clustering method
  - Evaluation of larger scale

~ Thank You ~